The Virus World, its evolution, evolution of antiviral defense, and the role of viruses in the evolution of cells

Eugene V. Koonin
National Center for Biotechnology Information, NIH, Bethesda

KITP, Santa Barbara, February 17, 2011
A **virus** is a small **infectious agent** that can replicate only inside the living cells of organisms. [http://en.wikipedia.org/wiki/Virus](http://en.wikipedia.org/wiki/Virus)

Viruses and virus-like agents **possess**:
- genomes
- very often –though not always – capsids that encase the genome

**but lack**:
- functional translation machinery
- membranes with transport/secretion systems

Viruses are the dominant entities in the biosphere—physically and genetically—as shown by viral metagenomics—virome studies.

1 cm³ of seawater contains $10^6$-$10^9$ virus particles


There are millions of diverse bacteriophage species in the water, soil, and gut


- Viruses are the most abundant biological entities in the biosphere: there are 10-100 virus particles per cell
- The pangenomes of viruses and cellular organisms have [at least] comparable complexities
High Frequency of Horizontal Gene Transfer in the Oceans

Lauren D. McDaniel,1,4 Elizabeth Young,1 Jennifer Delaney,4 Fabian Ruhnau,2 Kim B. Ritchie,3 John H. Paul1

Microbes rely on mutation and the processes of horizontal gene transfer (HGT; conjugation, transformation, and transduction) to acquire new traits. Gene transfer agents (GTAs) discovered in the purple nonsulfur bacterium Rhodobacter capsulatus (formerly Rhodopseudomonas capsulata) are host-encoded virulence elements that package random fragments of the host chromosome and are found in the genome of almost every sequenced member of the alpha-Proteobacteria order Rhodobacterales (1). To test whether GTAs are natural vectors of gene transfer, we grew nine strains of marine alpha-proteobacteria containing putative GTA cassettes (table S1) and screened them for the production of GTA-like particles.

Both Roseovarius nubilus ISM and the isolate Renanobacter mobilis 45A6 reproducibly produced putative GTA particles during stationary phase growth. We then generated genetically marked donor strains of R. nubilus and R. mobilis containing the transposon Tn5. GTA production in these marked donor strains was equivalent to that of the wild-type strains. To document gene transfer frequencies, we subjected wild-type strains or natural communities from a range of environments to treatment with donor strain GTAs and documented the rates of GTA-mediated gene transfer of kanamycin resistance (fig. S1). In the coral reef environment, spontaneous kanamycin resistance was $4.6 \times 10^{-4}$, whereas the GTA-mediated frequency was significantly higher at $2.5 \times 10^{-2}$ ($P = 0.028$, Student's $t$ test).

For this experiment, both spontaneous mutants and GTA treatments were examined for the presence of the Tn5 streptomycin kinase gene. A total of 47% of the GTA-treated viable colonies but none of the spontaneous revertants contained the gene. That 53% of the putative transductants did not contain the gene is not surprising because these may have contained only the kanamycin resistance gene (apdII) and not the flanking streptomycin kinase gene.

The recovery of the streptomycin kinase sequence, which is ~1000 base pairs (1 kb) from the active site of the kanamycin resistance gene, suggested that up to 1 kb of the central region of Tn5 was transferred. This is consistent with extracted DNA from the GTAs, which ranged from about 500 to 1600 bp in length (fig. S3). No spontaneous double antibiotic (kanamycin and streptomycin) resistance was detected, and the GTA-mediated frequency of $1.06 \times 10^{-4}$ was significantly higher ($P = 0.023$). The Tn5 streptomycin kinase sequence was recovered in 1 in 10 viable double antibiotic-resistant strains, suggesting that modifications, truncations, or rearrangements may have occurred, as in natural transformation (2).

Similar frequencies of transfer were observed among differing environments (Table 1), demonstrating that cultivated GTAs transduce natural communities of marine bacteria. The 16S ribosomal RNA sequences examined showed that the majority of natural GTA recipients were most similar to marine Flavobacterium or Flexibacter strains (table S2), consistent with the prior reports of abundant Flavobacterium in marine systems (3).

R. nubilus contains both a GTA and an inducible prophage (4). Transmission electron microscopy (TEM) demonstrated that R. nubilus-inducible prophage preparations contained tailed phage (4), whereas GTA particles were nontailed (fig. S2A), resembling the GTA of Silicibacter pomeroyi (5). In contrast to the GTA particles, the purified prophages of R. nubilus had no gene transfer activity. Additionally, maximal expression of the R. nubilus GTA terminase gene cooccurred with maximal GTA production (fig. S4). TEM of GTAs of R. mobilis revealed tailed viral particles (fig. S2B).

GTA dose, or multiplicity of infection (MOI), was linearly correlated with increased resistance to antibiotics (MOI range from 0.01 to 10, $R^2 = 0.9593$), which enabled extrapolations of gene transfer frequencies to natural systems (6).

GTAs from R. nubilus ISM show a wide host range and interspecies gene transfer under ecologically relevant conditions. Environmental gene transfer frequencies ranging from $6.7 \times 10^{-3}$ to $4.7 \times 10^{-1}$ (Table 1) are 1900 to 459 million times the frequency for transformation (2) and 650,000 to 31 million times the frequency of transduction previously measured in the marine environment (7). These results suggest a genomic flexibility in marine microorganisms that facilitates their adaptation to changing environmental conditions that facilitate their adaptation to changing environmental conditions.

References and Notes
Mean % of sequences with matches to major functional categories

Most of the viromes might not even consist of typical viruses but rather of pseudovirus particles that carry microbial genes (GTAs)

Kristensen, Mushegian, Dolja, Koonin Trends Microbiol 2010
Comparative genomics shows that viruses that cause human diseases belong to families that evolved hundreds of millions or even billions years ago:

Viruses accompany the evolving cellular life throughout its history and might even predate it
Some viruses are comparable to cellular life forms in size and genetic complexity!

The largest, most complex viruses: NCLDV (Nucleo-Cytoplasmic Large DNA viruses of eukaryotes) – this is where the smallpox virus belongs

*Mimivirus* genome (~1.2 Mbp, ~1,000 genes) is twice as large as that of *Mycoplasma genitalium* (580 kbp; ~500 genes)

The largest, most complex viruses: the Nucleocytoplasmic Large DNA Viruses (NCLDV) (this is where the smallpox virus AND the mimiviruses belong)

6 families of NCLDV...and counting

<table>
<thead>
<tr>
<th>Family</th>
<th>#</th>
<th>Size, kb</th>
<th>Hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>-poxviridae</td>
<td>26</td>
<td>[134-360]</td>
<td>vertebrates, insects</td>
</tr>
<tr>
<td>-asfarviridae</td>
<td>1</td>
<td>[170]</td>
<td>vertebrates, protists(?)</td>
</tr>
<tr>
<td>-iridoviridae</td>
<td>8</td>
<td>[103-212]</td>
<td>vertebrates, insects, protists(?)</td>
</tr>
<tr>
<td>-ascoviridae</td>
<td>4</td>
<td>[119-174]</td>
<td>insects</td>
</tr>
<tr>
<td>-phycodnaviridae</td>
<td>9</td>
<td>[155-407]</td>
<td>algae, haptophytes, stramenopiles</td>
</tr>
<tr>
<td>-mimiviridae</td>
<td>2</td>
<td>[1181-1200]</td>
<td>amoebozoa, algae(?)</td>
</tr>
<tr>
<td>-[Marseille virus]</td>
<td>1</td>
<td>[368]</td>
<td>amoebozoa – new family?</td>
</tr>
</tbody>
</table>

The case for the monophyly (common origin) of NCLDV

9 universally conserved hallmark genes (vaccinia gene names):
- primase (D5-N)
- helicase (D5-C)
- DNA polymerase (E9)
- packaging ATPase (A32)
- Major capsid protein (D13, non-capsid in poxviruses)
- Thiol-oxidoreductase (E10)
- Helicase (D6, D11)
- S/T protein kinase (F10)
- Transcription factor VLTF2 (A1)

47 genes mapped to the last common ancestor of NCLDV by maximum likelihood – all main functional systems represented

Phylogeny of NCLDV based on concatenation of 4 universal genes: primase-helicase, DNA polymerase, packaging ATPase, VLTF2 transcription factor.

HOST

Animals

+ diverse protists

Amoebozoa, Algae, animals(?)

Chlorophytes, Haptophytes, stramenopiles

Phylogeny of NCLDV most likely antedates divergence of eukaryotic supergroups

• Alternative/addition: extensive horizontal transfer of viruses

Asfarviridae

Animals, haptophytes

Marseillevirus

Mimiviridae

Phycodnaviridae

Poxviridae

Ascoviridae & Iridoviridae

• Divergence of NCLDV most likely antedates divergence of eukaryotic supergroups

Boyer, Yutin et al PNAS 2009
Koonin, Yutin, Intervirology 2010

HOST

Animals

+ diverse protists

Amoebozoa, Algae, animals(?)

Chlorophytes, Haptophytes, stramenopiles

Phylogeny of NCLDV most likely antedates divergence of eukaryotic supergroups

• Alternative/addition: extensive horizontal transfer of viruses

Asfarviridae

Animals, haptophytes

Marseillevirus

Mimiviridae

Phycodnaviridae

Poxviridae

Ascoviridae & Iridoviridae

• Divergence of NCLDV most likely antedates divergence of eukaryotic supergroups

Boyer, Yutin et al PNAS 2009
Koonin, Yutin, Intervirology 2010
Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms.

Giant viruses such as Mimivirus isolated from amoeba found in aquatic habitats show biological sophistication comparable to that of simple cellular life forms and seem to evolve by similar mechanisms, including extensive gene duplication and horizontal gene transfer (HGT), possibly in part through a viral parasite, the virophage. We report here the isolation of "Marseille" virus, a previously uncharacterized giant virus of amoeba. The virions of Marseillevirus encompass a 368-kb genome, a minimum of 49 proteins, and some messenger RNAs. Phylogenetic analysis of core genes indicates that Marseillevirus is the prototype of a family of nucleocytoplasmic large DNA viruses (NCLDV) of eukaryotes. The genome repertoire of the virus is composed of typical NCLDV core genes and genes apparently obtained from eukaryotic hosts and their parasites or symbionts, both bacterial and viral. We propose that amoebae are "melting pots" of microbial evolution where diverse forms emerge, including giant viruses with complex gene repertoires of various origins.
The mosaic composition of the Marseille virus genome
Gene content tree: intervirus gene transfer

African swine fever virus
- *Amsacta moorei* entomopoxvirus
  - *Melanoplus sanguinalis* entomopoxvirus
    - Canarypox virus
    - Fowlpox virus
    - Bovine papular stomatitis virus
    - Orf virus
    - Crocodilepox virus
    - Molluscum contagiosum virus
    - Vaccinia virus
    - Variola virus
    - Myxoma virus
    - Rabbit fibroma virus
    - Goatpox virus Pello
    - Sheeppox virus
    - Lumpy skin disease virus
    - Tzanapox virus
    - Yaba-like disease virus
    - Yaba monkey tumor virus
    - Deerpox virus
    - Swinepox virus

Marseille virus
- *Acanthamoeba polyphaga* mimivirus
  - *Acanthamoeba polyphaga* mamavirus
    - *Acanthamoeba surface* Chlorella virus 1
      - *Paramecium bursaria* Chlorella virus FR483
        - *Paramecium bursaria* chlorella virus MT325
    - *Paramecium bursaria* Chlorella virus AR158
    - *Paramecium bursaria* Chlorella virus NY24A
    - *Paramecium bursaria* Chlorella virus 1
    - *Ostreococcus* virus OsV5
      - *Ectocarpus siliculosus* virus 1
      - Feldmannia species virus
    - *Emiliania huxleyi* virus 86

- Lymphocytic choriomeningitis virus
  - Lymphocytic disease virus - isolate China
    - *Anthystoma tigrinum* virus
      - Frog virus 3
      - Singapore grouper iridovirus
    - Infectious spleen and kidney necrosis virus
      - *Aedes taeniorhynchus* iridescent virus
      - Invertebrate iridescent virus 6
    - *Trichoplax adhaerens* ascovirus 2c7
      - *Spodoptera frugiperda* ascovirus 1a
      - *Heliotis virescens* ascovirus 3e
Amoeba as a melting pot for HGT between viruses and bacterial endosymbionts
There are really weird creatures out there: Some NCLDV host their own parasites

Chimeric origin of the virophage genome

La Scola et al. Nature, 2008; 455:100-4
Hypothesis: origin of the NCLDV (and other viruses of Eukaryotes) in the second melting pot of virus evolution – eukaryogenesis

Phage scaffold (virus hallmark genes)

Eukaryotic additions/displacements

A host: Archaeon with plasmids, group I introns, and viruses

A symbiont: α-Proteobacterium with plasmids, group I & II introns, retrons, and phages

Koonin, Yutin, Intervirology 2010
Some of the smallest viruses (this is where poliovirus belongs): The Big Bang of picorna-like virus evolution

Picorna-like viruses possess diverse arrays of hallmark and unique genes

Marine eukaryotic plankton carries a wealth of positive-strand RNA viruses: nearly all belong to the picorna-like superfamily.

- **HcRNAV**
  - Size: 4.4 kb
  - Genome structure: S-Pro, Pol, VP

- **RsRNAV**
  - Size: 8.9 kb
  - Genome structure: Hel3, Pol, VP 1-3

- **HaRNAV**
  - Size: 8.6 kb
  - Genome structure: Hel3, Pol, VP2, VP3, VP1

- **SssRNAV**
  - Size: 9.0 kb
  - Genome structure: Hel3, C-Pro, Pol, VP 1-3

References:
- Nagasaki et al. (2005) *Appl. Env. Microbiol.* **71**:8888
- Culley et al. (2006) *Science* **312**:1795
The amended Picornavirus-like superfamily includes 14 recognized viral families, 4 floating genera and 15 unclassified positive-strand and double-strand RNA viruses that infect hosts from 4 of the 5 eukaryotic supergroups - 6 distinct clades from RdRp phylogeny - diverse genome layouts.
View from the viral side

6 major clades of picorna-like viruses - 5 infect eukaryotes from different supergroups
The 5 supergroups of eukaryotes and their picorna-like viruses

Complementary view from the host side

4 of the 5 eukaryotic supergroups host diverse picorna-like viruses (no one checked the 5th)
Likely origins of the signature genes of the picorna-like superfamily were inferred using signature PSSMs and PSI-BLAST searches.
Hallmark genes of picorna-like viruses have distinct prokaryotic origins

Radiation of major viral clades occurred in a “Big Bang” during eukaryogenesis and antedates the divergence of eukaryotic supergroups - the viruses then “sampled” the hosts.
Conclusions on the picorna and NCLDV stories:

**Big Bangs of virus evolution**

- Phylogenomic analysis of the picorna-like superfamily of eukaryotic RNA viruses indicates that the major lineages within this superfamily diverged prior to the divergence of the eukaryotic supergroups
- **Big Bang** – an explosive early phase of viral evolution
- Most likely, the same pattern holds for other major groups of viruses as illustrated by the evolutionary study of NCLDV, a completely different group of viruses
- The **Big Bangs** of eukaryotic virus evolution occurred concomitantly with a similar rapid phase of host evolution and could be a manifestation of a general model of major evolutionary transitions

**Diversity of viral genetic cycles versus the uniform genetic cycle of all cellular organisms:** Viruses are the biosphere’s laboratory of genomic strategies

### Genetic cycles of RNA viruses

<table>
<thead>
<tr>
<th>Class</th>
<th>Replication cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Positive-strand RNA</strong> 3-30 kb</td>
<td><img src="image" alt="Diagram of positive-strand RNA replication cycle" /></td>
</tr>
<tr>
<td><strong>2. Double-strand RNA 4-25 kb</strong></td>
<td><img src="image" alt="Diagram of double-strand RNA replication cycle" /></td>
</tr>
<tr>
<td><strong>3. Negative-strand RNA 11-20 kb</strong></td>
<td><img src="image" alt="Diagram of negative-strand RNA replication cycle" /></td>
</tr>
</tbody>
</table>
Genetic cycles of retroid viruses and retroelements

<table>
<thead>
<tr>
<th>Class</th>
<th>Replication cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Retroid RNA viruses, 7-12 kb</td>
<td>RT Tr, E T ± Tr, RT, CP</td>
</tr>
<tr>
<td>5. Retroid DNA viruses, elements 2-10 kb</td>
<td>Tr, Tr, E, RT</td>
</tr>
</tbody>
</table>

Genetic cycles of DNA viruses and plasmids

<table>
<thead>
<tr>
<th>Class</th>
<th>Replication cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. ssDNA viruses, plasmids, 2-11 kb</td>
<td>RCR Tr, RCE</td>
</tr>
<tr>
<td>7. dsDNA viruses, plasmids, 5-1,200 kb</td>
<td>Tr, Pr-Pol, R, E</td>
</tr>
</tbody>
</table>
Natural history of viral genes: a one-page summary of viral comparative genomics

I. Genes with readily detectable homologs from cellular life forms:
1. Genes with closely related homologs from cellular organisms (typically, the host of the given virus) present in a narrow group of viruses
2. Genes that are conserved within a virus lineage or even several lineages and have moderately close cellular homologs
   
   Origin: relatively recent (1) or ancient (2) acquisition from host

II. Virus-specific genes
3. ORFans, i.e., genes without detectable homologs except, possibly, in closely related viruses
4. Virus-specific genes that are conserved within a virus lineage

   Acquisition from host but with rapid divergence from ancestor once within viral genomes?

III. Viral hallmark genes
5. Genes shared by many diverse virus lineages, with only very distant homologs in cellular organisms
Contributions of different classes of viral genes to the genomes of different classes of viruses: strong dependence on genome size
Natural history of viral genes: Viral Hallmark Genes

Shared by many diverse groups of viruses: from the smallest RNA viruses to the giant DNA viruses

Strong support for monophly of all viral members of the respective gene families

Only distant homologs in cellular organisms

Play major roles in genome replication, packaging and assembly

Can be viewed as signatures of the ‘virus state’
Protein products of viral hallmark genes

1. Jelly-roll capsid protein
2. Superfamily 3 helicase
3. RNA-dependent RNA polymerase and Reverse transcriptase
4. Rolling circle replication initiation endonuclease
5. Viral archaeo-eukaryotic DNA primase
6. UL9-like superfamily 2 helicase
7. Packaging ATPase of the FtsK family
8. ATPase subunit of terminase
Viral hallmark genes are:
- present in a huge diversity of viruses and other selfish elements
- represented only by remote homologs in cellular life forms

The hallmark genes AND, by implication, the major lineages of modern viruses (at least, viruses of prokaryotes) descend directly from a primordial gene pool

-synergy with the diversity of genomic strategies

A crucial corollary: If viruses come directly from a primordial gene pool, then, origin of viruses is inextricably linked to the origin of cells
Competing concepts of the origin of viruses

Cell degeneration

- Cell
  - Small parasitic cell
  - Very small virus

Escaped genes

- Chromosome
  - mRNA
  - Plasmid
  - Virus

Primordial genetic systems

- Pre-cellular life forms
  - RNA
  - Virus
  - DNA
  - Virus
  - Cells
Origin and evolution of virus-like genetic elements in the pre-cellular era

Koonin, Martin, TIG, 2005
Koonin, Senkevich, Dolja.
Biol. Direct. 2006, 1:29
Koonin, Ann NY Acad Sci, 2009
• Viruses and virus-like genetic elements are not “just” pathogens: they are dominant entities in the biosphere
• Emergence of virus-like parasites is inevitable in any replicating system
• In the pre-cellular epoch, the genetic elements that later became viral and cellular genomes comprised a single pool in which they mixed, matched, and evolved new, increasingly complex gene ensembles
• Different replication strategies including RNA replication, reverse transcription, and DNA replication evolved already in the primordial genetic pool
• With the emergence of prokaryotic cells, a distinct pool of viral genes formed that retained its identity ever since as evidenced by the extant distribution of viral hallmark genes: “virus world” or the virosphere
• The emergence of the eukaryotic cell was a second melting pot of virus evolution, from which viruses of eukaryotes originated via recombination of genes from prokaryote viruses, retroelements, and the evolving eukaryotic host
• Viruses make essential contributions to the evolution of the genomes of cellular life forms, in particular, as vehicles of HGT: GTAs, transducing phages
Evolution of antivirus defense systems

- CRISPR/Cas system of adaptive immunity in prokaryotes
- A case for Lamarckian evolution
- The perpetual virus-host arms race
CRISPR repeats and Cas genes

CRISPR: Clustered, Regularly interspaced short palindromic repeats
Cas: CRISPR-associated (genes)

A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.

During a systematic analysis of conserved gene context in prokaryotic genomes, a previously undetected, complex, partially conserved neighborhood consisting of more than 20 genes was discovered in most Archaea and some bacteria, including the hyperthermophiles Thermotoga maritima and Aquifex aeolicus. The gene composition and gene order in this neighborhood vary greatly between species, but all versions have a stable, conserved core that consists of five genes. One of the core genes encodes a predicted DNA helicase, often fused to a predicted HD-superfamily hydrolase, and another encodes a RecB family exonuclease; three core genes remain uncharacterized, but one of these might encode a nuclease of a new family.

The functional features of the proteins encoded in this neighborhood suggest that they comprise a previously undetected DNA repair system, which, to our knowledge, is the first repair system largely specific for thermophiles to be identified.
Protein components of the system: an update with unification of many diverse families ~25 families altogether

<table>
<thead>
<tr>
<th>Family Subfamily</th>
<th>Phyletic distribution</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG1518</td>
<td>COG1518 (cas1)</td>
<td>All</td>
</tr>
<tr>
<td>COG1343</td>
<td>COG1343 (cas2), COG3512, ygbF-like; MTH324-like; y1723_N-like;</td>
<td>All</td>
</tr>
<tr>
<td>COG1203</td>
<td>COG1203 (cas3)</td>
<td>All</td>
</tr>
<tr>
<td>RecB-like nuclease</td>
<td>COG1468 (cas4), COG4343</td>
<td>All</td>
</tr>
<tr>
<td>RAMP: Repair-associated mysterious protein</td>
<td>COG1688 (cas5), COG1769, COG1583,COG1567, COG1336,COG1367, COG1604,COG1337, COG1332,COG5551, BH0337-like, MJ0978-like, YgcH-like, y1726-like, y1727-like</td>
<td>All</td>
</tr>
<tr>
<td>COG1857</td>
<td>COG1857, COG3649, YgcJ-like, y1725-like</td>
<td>All</td>
</tr>
<tr>
<td>HD-like nuclease</td>
<td>COG1203 (N-terminus), COG2254</td>
<td>All</td>
</tr>
<tr>
<td>BH0338</td>
<td>BH0338-like, MTH1090-like</td>
<td>All, mostly archaea and FIRM</td>
</tr>
<tr>
<td>COG1353, predicted polymerase</td>
<td>COG1353</td>
<td>Most archaea, some bacteria</td>
</tr>
</tbody>
</table>

...and ~15 other, less common proteins

Makarova et al. BD 2006
CRISPR: clustered regularly interspaced short palindromic repeats

Identification of genes that are associated with DNA repeats in prokaryotes.

Jansen R, Embden JD, Gaastra W, Schouls LM.

Department of Infectious Diseases and Immunology, Bacteriology Division, Veterinary Faculty, Utrecht University, Yalelaan 1, 3584 CL Utrecht, The Netherlands. R.jansen@vet.uu.nl

Using in silico analysis we studied a novel family of repetitive DNA sequences that is present among both domains of the prokaryotes (Archaea and Bacteria), but absent from eukaryotes or viruses. This family is characterized by direct repeats, varying in size from 21 to 37 bp, interspaced by similarly sized non-repetitive sequences. To appreciate their characteristi-c structure, we will refer to this family as the clustered regularly interspaced short palindromic repeats (CRISPR). In most species with two or more CRISPR loci, these loci were flanked on one side by a common leader sequence of 300-500 b. The direct repeats and the leader sequences were conserved within a species, but dissimilar between species. The presence of multiple chromosomal CRISPR loci suggests that CRISPRs are mobile elements. Four CRISPR-associated (cas) genes were identified in CRISPR-containing prokaryotes that were absent from CRISPR-negative prokaryotes. The cas genes were invariably located adjacent to a CRISPR locus, indicating that the cas genes and CRISPR loci have a functional relationship. The cas3 gene showed motifs characteristic for helicases of the superfamily 2, and the cas4 gene showed motifs of the RecB family of exonucleases, suggesting that these genes are involved in DNA metabolism or gene expression. The spatial coherence of CRISPR and cas genes may stimulate new research on the genesis and biological role of these repeats and genes.
“The common structural characteristics of CRISPR loci are: (i) the presence of multiple short direct repeats, which show no or very little sequence variation within a given locus; (ii) the presence of non-repetitive spacer sequences between the repeats of similar size; (iii) the presence of a common leader sequence of a few hundred basepairs in most species harbouring multiple CRISPR loci; (iv) the absence of long open reading frames within the locus; and (v) the presence of the cas1 gene accompanied by the cas2, cas3 or cas4 genes in CRISPR-containing species.”

The cas genes are our “repair” system!
paragon of prokaryotic genome plasticity:
- extensive gene shuffling
- evidence of multiple horizontal transfers
- widespread gene loss
**RAMP (Repeat-Associated Mysterious Proteins) superfamily:** numerous families of Cas proteins, extreme sequence diversity

<table>
<thead>
<tr>
<th>MOTIFs</th>
<th>specific</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>COGs 1336,1367, 1604,1337,1332</td>
<td>h.h...h.h.k</td>
<td>-</td>
<td>hG+p.khh+h.h</td>
<td>hG+p</td>
<td>hD</td>
<td>hG+p.c.c.s.c.r</td>
</tr>
<tr>
<td>y1726-like</td>
<td>hlpEKVRh</td>
<td>hhp2c.p2 Rh</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>COG1851</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>BH0337-like</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>COG1567</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>COG1769</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>COG1688 (Cas5)</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>COGs 1583,5551</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>YgCH-like</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>YgCH-like</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
<tr>
<td>MJ0978-like</td>
<td>hlp2c.p2</td>
<td>hhp2c.p2</td>
<td>hhp2c.p2</td>
<td>hlp2</td>
<td>hlp2</td>
<td>hlp2</td>
</tr>
</tbody>
</table>

**Ferredoxin fold/RNA-recognition motif (RRM)**
CRISPR show extreme diversity and complex clustering

The sequence similarity space of CRISPR repeats visualized with the BioLayout program [26]. Dots denote individual repeat sequences; connecting lines represent Smith-Waterman similarities, such that closer dots represent more similar sequences. Dot colors denote cluster association as derived from MCL clustering. The 12 largest clusters are indicated by circles together with their sequence logos, coarse phylogenetic composition, and sample secondary structures where applicable. Kunin et al. Genome Biology 2007 8:R61 doi:10.1186/gb-2007-8-4-r61
A large subset of CRISPRs is conserved among diverse species, even between archaean and bacteria, suggesting that CRISPRs are horizontally transferred together with cas genes.
Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. 
J Mol Evol. 2005 Feb;60(2):174-82

Here we show that CRISPR spacers derive from preexisting sequences, either chromosomal or within transmissible genetic elements such as bacteriophages and conjugative plasmids.

The transcription of the CRISPR loci (Tang et al. 2002) suggests that such activity could be executed by CRISPR-RNA molecules, acting as regulatory RNA that specifically recognizes the target through the homologous RNA-spacer sequence, similarly to the eukaryotic interference RNA.
Biology Direct

Research

A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action

Kira S Makarova¹, Nick V Grishin², Svetlana A Shabalina¹, Yuri I Wolf¹ and Eugene V Koonin*¹

Address: ¹National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and ²Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Email: Kira S Makarova - makarova@ncbi.nlm.nih.gov; Nick V Grishin - grishin@chop.swmed.edu; Svetlana A Shabalina - shabalina@ncbi.nlm.nih.gov; Yuri I Wolf - wolf@ncbi.nlm.nih.gov; Eugene V Koonin* - koonin@ncbi.nlm.nih.gov

* Corresponding author

Published: 16 March 2006
Received: 08 February 2006
Accepted: 16 March 2006


This article is available from: http://www.biology-direct.com/content/1/1/7

© 2006 Makarova et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract
Hypothesis

CRISPR/Cas

• is a prokaryotic immunity system that functions on the RNAi principle

• integrates short fragments of essential phage/plasmid genes into CRISPRs

• When expressed, these fragments (psiRNA – after prokaryotic siRNA) silence the target gene and make the organism immune to the respective agent

• contains all or most of the protein activities involved in these processes

• Some of the Cas proteins are functional analogs of the eukaryotic proteins involved in RNAi, in particular, components of RISC (RNA-Induced Silencing Complex), and form prokaryotic analogs of RISC (pRISCs)
The basic scheme of CASS functioning

- Specific transcription factor (COG1517?)
- Cellular RNApol
- Transcription
- (protein-guided) RNA folding
- Pre-psiRNA 75-100 nt
- RNA processing (fast?)
- Pre-psiRNA 75-100 nt
- RNA processing (slow?)
- psiRNA 25-45 nt
- Annealing to RNA target
- p-RISC
- Target RNA cleavage
- Plasmid or phage mRNA

Annotations:
- p-dicer (Helicase+HD-Hydrolase) (COG1203? COG2254?)
- p-dicer
- RAMP
- p-slicer (COG1468? COG4343? COG1857?)
Variant of CASS functioning with polymerase/psi-RNA amplification (mostly, in thermophiles, but also in Mycobacteria)

- RAMP
- pmRNA
- psiRNA 25-45 nt
- RAMP-RNA complex (unstable)
- plasmid or phage mRNA
- CASS RNApol (COG1353)
- Primer elongation
- long dsRNA (stable)
- Duplex degradation
- RAMP binding
- Amplified annealing
- Cycle continues
New psiRNA generation

Random RNA recombination and reverse transcription

OR

Reverse transcription with random copy choice

dsDNA with CRISPRs and target-derived spacers

Homologous recombination with genomic CRISPR region

genomic DNA with new target-derived spacer
Key validation:

CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007 Mar 23;315(5819):1709-12

Clustered regularly interspaced short palindromic repeats (CRISPR) are a distinctive feature of the genomes of most Bacteria and Archaea and are thought to be involved in resistance to bacteriophages. We found that, after viral challenge, bacteria integrated new spacers derived from phage genomic sequences. Removal or addition of particular spacers modified the phage-resistance phenotype of the cell. Thus, CRISPR, together with associated cas genes, provided resistance against phages, and resistance specificity is determined by spacer-phage sequence similarity.
Phage-specific inserts confer resistance that is highly sequence-specific: a single substitution (SNP) reverts to sensitivity.
The spacers worked only when inserted between CRISPR.
Resistance required COG3513 (cas5), a predicted nuclease.
Inserts from phage-resistant mutants were homologous to regions scattered over the phage genomes.
(1) Adaptation

invading virus / plasmid

Cas3

leader

CRISPR

host

(2) Expression & Processing

Cas3 cas2 cas1 cas5

Cascade

host

(3) Interference

Cas3

invading virus / plasmid

Van der Oost et al. TIBS 2009
The three types of CRISPR/Cas systems and their signature genes

Cas1  Cas2

Cas6

Makarova et al, NRM submitted
Experimental data on CRISPR/Cas systems

**Streptococcus thermophilus**
Barrangou, R. et al. Science 315, 1709-12 (2007);

**Staphylococcus epidermidis**

**Pyrococcus furiosus**

**E. coli**

**Pseudomonas aeruginosa**
Phylogeny of Cas1 and the 3 types of CRISPR/Cas systems

228 representative Cas1 sequences, selected out of 643 Cas1 proteins in 442 genomes

Makarova et al., NRM submitted
CRISPR/CAS systems in 703 selected complete genomes of archaea and bacteria

- Cas1 is present in 310 (44%) genomes
- ~90% archaea but only ~35% of bacteria
- Type I is present in 42% genomes; Type II – 9%; Type III – 20%;
- Two or three systems of different types are present in 128 (20%) genomes

Makarova et al, in preparation
Modular evolution of the 3 types of CRISPR/Cas

Makarova et al, in preparation
A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair.


Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and the associated proteins (Cas) comprise a system of adaptive immunity against viruses and plasmids in prokaryotes. Cas1 is a CRISPR-associated protein that is common to all CRISPR-containing prokaryotes but its function remains obscure. Here we show that the purified Cas1 protein of Escherichia coli (YgbT) exhibits nuclease activity against single-stranded and branched DNAs including Holliday junctions, replication forks and 5'-flaps. The crystal structure of YgbT and site-directed mutagenesis have revealed the potential active site. Genome-wide screens show that YgbT physically and genetically interacts with key components of DNA repair systems, including recB, recC and ruvB. Consistent with these findings, the ygbT deletion strain showed increased sensitivity to DNA damage and impaired chromosomal segregation. Similar phenotypes were observed in strains with deletion of CRISPR clusters, suggesting that the function of YgbT in repair involves interaction with the CRISPRs. These results show that YgbT belongs to a novel, structurally distinct family of nucleases acting on branched DNAs and suggest that, in addition to antiviral immunity, at least some components of the CRISPR-Cas system have a function in DNA repair.
The 3 major modalities of evolution

**Lamarck**
- Environmental factors
- Mutation-directing mechanism
- Beneficial mutations
- Adapted organism

**Darwin**
- Environmental factors
- Random mutagenesis
- Selection
- Beneficial mutations fixed by selection; adapted organism

**Wright**
- Random mutagenesis
- Random fixation
- Beneficial mutations fixed by chance; adapted organism

Koonin, Wolf, Biol. Direct 2009
CRISPR/Cas as a bona fide Lamarckian system

Koonin, Wolf, Biol. Direct 2009
<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Biological role/function</th>
<th>Phyletic spread</th>
<th>Lamarckian criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Genomic changes caused by environmental factor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Changes are specific to relevant genomic loci</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Changes provide adaptation to the causative factor</td>
</tr>
<tr>
<td><strong>Bona fide Lamarckian</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRISPR/Cas</td>
<td>Defense against viruses and other mobile elements</td>
<td>Most of the Archaea and many bacteria</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>piRNA</td>
<td>Defense against transposable elements in germline</td>
<td>Animals</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>HGT (specific cases)</td>
<td>Adaptation to new environment, stress response, resistance</td>
<td>Archaea, bacteria, unicellular eukaryotes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Quasi-Lamarckian</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HGT (general phenomenon)</td>
<td>Diverse innovations</td>
<td>Archaea, bacteria, unicellular eukaryotes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes/no</td>
</tr>
<tr>
<td>Stress-induced mutagenesis</td>
<td>Stress response/resistance/adaptation to new conditions</td>
<td>Ubiquitous</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>No or partially</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes (but general evolvability enhanced as well)</td>
</tr>
</tbody>
</table>
Stress as a gauge of evolutionary modality

Lamarckian modality

stress level

Darwinian modality

Koonin, Wolf, Biol. Direct 2009
• Evolution of parasites is intrinsic to any replicator system

• Defense systems, in particular, those based on the RNAi principle, appeared concomitantly with cells and coevolved with cells and viruses ever since

• Defense systems occupy a substantial fraction of the genomes in all cellular life forms

• Perennial arms race between parasites and hosts is one of the principal factors of evolution